Accepted for presentation at the NIH Research Festival 2006.
draft material -- please do not cite

Integrating biomedical knowledge: Experience with Entrez Gene

Olivier Bodenreider, Satya Sahoo, Kelly Zeng

Modern biomedical research is increasingly supported by information technologies. Biologists and physicians rely not only on the biomedical literature (e.g., MEDLINE), but also on the many knowledge bases available online (e.g., through the National Center for Biotechnology Information's Entrez portal). While these resources are undeniably valuable to humans, most of them are text-based and heterogeneous, and cannot be easily processed by computers.

The Biomedical Knowledge Repository under development at the National Library of Medicine addresses these limitations. It can be understood as a specialized version of the Semantic Web. It consists of an extensive collection of assertions (i.e. concept-relationship-concept triples), represented in a common format, processable by computers. Logical reasoners extend the capabilities of the repository by inferring new knowledge.

We converted NCBI's gene information resource Entrez Gene from its XML format to RDF, the Resource Description Framework. This transformation is not simply syntactic, but also semantic and was achieved through rules created manually and expressed in an XSLT (Extensible Stylesheet Language Transformation).

Once converted to RDF, Entrez Gene can be integrated seamlessly with other resources such as MeSH. Hierarchies in MeSH can then be exploited to query genes, enabling researchers to formulate queries such as "Find all genes involved with neurodegenerative diseases".